Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS

Miloš Ercegovčević

Joint Learning of Syntax and Semantics

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: asst. prof. RNDr. Ondřej Bojar, Ph.D.
Study programme: Informatics
Specialization: Mathematical Linguistics and Informatics

Prague 2012

Making a statement of gratitude of acknowledgement at a single point in this continuum of ours is an ungrateful endeavor. We risk of being momentarily influenced and deceived by our various limitations. Thus I would like to thank all the great people that had a privilege too meet, hear, interact or even being inspired by. As we jump between our misconceptions we sooner or later became aware of them and greatly appreciate. Thank you!

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... date .... signature of the author

Název práce: Společná učení syntaxí a sémantiky

Autor: Miloš Ercegovčević

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: asst. prof. (odb. as.) RNDr. Ondřej Bojar, Ph.D.

Abstrakt: Tato diplomová práce se zabývá problémem vzdělávání různé úrovně abstrakce jazykové znalosti. Hlavní důraz je kladen na učení latentní sémantické informace zastoupená rámu tříd, slovesných tříd a role vložky podle průzkumu nedávné úspěchy v Bayesian modelování se skrytou proměnné. Dále je v blízkosti spojka syntaxe a sémantiky zachycen v společný model, který také obsahuje cenné lexikální informace. Výsledek je jazykově nezávislý, rys-menší model sémantické informace se výkon srovnatelný se současným stavem techniky.

Klíčová slova: sémantika, syntax, společně učení, latentní proměnné, jazyk–nezávislé

Title: Joint Learning of Syntax and Semantics

Author: Miloš Ercegovčević

Department: Institute of Formal and Applied Linguistics

Supervisor: asst. prof. (odb. as.) RNDr. Ondřej Bojar, Ph.D.

Abstract: This master thesis addresses the problem of learning varying levels of abstraction of linguistic knowledge. The main focus is on learning latent semantic information as represented by frame classes, verb classes and role fillers by exploring recent successes in Bayesian modeling with hidden variables. Furthermore, close coupling of syntax and semantics is captured in the joint model which also incorporates valuable lexical information. The result is a language-independent, feature-less model of semantic information with a performance comparable to the current state of the art.

Keywords: semantics, syntax, joint learning, latent variables, language–independent

Contents

1. Introduction

1.1 Semantics in Natural Language Processing

Empiricist methods are dominant in the field of Computational Linguistics, ranging from simple tasks, such as part-of-speech speech tagging, chunking, named entity recognition to more complex tasks like syntactic parsing, speech recognition or machine translation. After huge improvements in stochastic parsing of natural languages, the field has begun to impose tasks that involve a higher level of abstraction such as ***semantic parsing***, going toward semantic understanding. Even though characterizing 'who' did 'what' to 'whom,' 'where,' 'when,' and 'how' might not solve the long-reaching goals of Artificial Intelligence, the task had some successful application domains such as information extraction, question answering, summarization and machine translation. Starting from purely supervised approaches to semantic parsing, recent research also shows quite promising results in unsupervised methods. However, several fundamental questions remain even in this shallow form of semantic parsing. Namely, levels of abstraction have been de facto imposed by annotated corpora such as FrameNet and Propbank, which in turn have been shown to be very limiting in their out-of-domain performance. Furthermore, state-of-the-art performance is achieved by feature engineering, which is a very tedious and time-consuming task, one which usually does not scale neither domain-wise nor language-wise. Going toward the highly ambitious goal of defining direct correspondences between natural languages on the semantic level will undoubtedly have to tackle those problems. While unsupervised approaches have tried to resolve some of the problems above, their performance in even simple tasks such as part-of-speech tagging is questionable. This master thesis will try to tackle learning of ***latent semantic representations***in a supervised setting with ***varying levels of abstraction***, by jointly learning syntactic and semantic dependencies and evaluating them on data provided by CoNLL09 shared task.

1.2 Modeling Semantics

The picture that had emerged in previous years is that corpora annotation annotations driven by linguistic consensus are not the most representative for self-elaboration and prediction. Various approaches have been devised to tackle this ***un–represe ntativeness***, mostly employing discriminative machine learning approaches. However, generative approaches in general have shown to be better on a lower scale and have a nice convenient property that they can handle missing, incomplete data and can incorporate latent variables. Exploiting recent successes in Bayesian modeling with hidden variables, in this master thesis we use a generative latent variable model to tackle ***joint learning of syntax and semantics***. Assuming exact number of semantic roles and predicate fillers has been assumed so far in context of ProbBank and learning them has not been attempted. However, starting from the FrameNet intuition of ***the hierarchy of the semantic frames*** we threat learning of semantic abstraction as hidden information, which should maximize the likelihood of data. ***Role fillers*** have been modeled so far as verb-specific and even corpus annotations have assumed them to be such, as in PropBank, simple investigation shows that they share a lot of lexical or syntactic similarity. Furthermore, role fillers generalize among themselves with inhibition of different lexical information; we model such a behavior with latent lexical categories which can serve as word class information. Previous years have also shown that splitting ***syntactic dependencies*** can be helpful for parsing as shown by specification of some dependencies in CoNLL08 task. We handle them as well with latent variables. The de facto linguistic background of jointly learning syntax and semantics is driven by the ***linguistic theory of linking***. Linking theory implies that syntactic behavior can be determined from the underlying semantics. We model linking alternations jointly in our model with latent frames, semantics, syntactic dependencies and lexical categories. Such an approach can be seen as both supervised, as the backbone structures are provided to us, and unsupervised because the model softly clusters the labeled variables into statistically dependent groups, resulting in what one may call ***semi–supervised learning***.

1.3 Road Map

This chapter gives a critical introduction to recent research practices in the task of semantic role labelling. Chapter 2 introduces necessary theoretical and practical properties on treating semantics in the field of Computational Linguistics. We explore two most common annotation schema PropBank and FrameNet. In the chapter 4 we briefly point to computational treatment of supervised and unsupervised approaches on the task of semantic role labeller. Paragraph 5 introduces our key scientific framework of latent probabilistic context–free grammars for modelling semantics. Finally in paragraph 6 we tackle the problem both with modelling and technical prospective. We present our latent variable model that without any features automatically learns appropriate representations and perform well on the task of interest. Paragraphs 6 and 7 comment on empirical results and future work respectively.

2. Semantic theory and Computational Resources

2.1 From linguistic theory to computational practice

Semantic analysis of sentence-level utterances deals with characterizing of events and their participants. The event is activated by the event invoker that characterizes 'what' took place, which further specifies the 'who' and 'whom' has the processes undergone and some general properties like 'where' or 'when'[9]. The event by itself is a carrier of the information and is most usually represented by a predicate, while the participants and properties define roles with respect to the predicate.

Consider for example the following sentence[1]:

- [The girl on the swing]$_{\text{Agent}}$ [whispered]$_{\text{Pred}}$ to [white boy beside her]$_{\text{Recipient}}$.

Defining this example from same level of abstraction, we can say that the Conversation event is invoked by the predicate 'whispered' and that the participant 'the girl on the swing' is the agent of the event while 'the boy beside her' is the patient.

The theory of semantic roles goes far as thousands of years in Panini's Karaka theory. The whole ***spectrum of generality*** of the roles has been defined in theory as well as in practice[3]. At one end of the spectrum, there are specific roles such as FromAirport, ToAirport or Depart that found useful application in natural language understanding specifically in dialog systems. On the other end, there are more coarse-grained roles that can be merged down to as few as two roles (Proto-agent and Proto-patient) to several roles such as Fillmore's list of nine: Agent, Experiencer, Instrument, Object, Source, Goal, Location, Time, and Path. The more general roles have been proposed by the linguists who are more interested in describing generalizations across syntactic realizations of their arguments as driven linguistic theory of linking. On the other hand, computer scientists have been proposing more specific roles as they are more interested in details of the realization of the arguments.

2.1.1 Linking Theory

Linking theory argues that the alternation behavior of the verb as described by the syntactic frames is a direct reflection of the underlying semantics[8]. The theory introduces Levin verb classes, which are defined by the syntactic frames which respectively constrain allowable arguments of semantics. Thus a verb class is defined as the possibility of a particular verb to occur in pairs of syntactic frames. It is further argued that the syntactic frames are meaning-preserving

[1] Example taken from [9].

and that classes tend to share some of the semantic behavior; the principle is called ***diathesis alternations***.

For example, let us consider alternations of break verbs 'break', 'shatter' and 'smash'. All of them can be characterized in their ability to occur in the middle construction like in[2]:

- Glass breaks/shatters/smashes easily.

Now consider the verb 'cut' which is very similar to the verbs above and also tends to occur in the middle construction, like in[2]:

- John cut the bread.

However 'cut' cannot occur in the intransitive construction like in 'The bread cut', while 'The window broke' is very plausible. On the other hand 'cut' can occur in the conative like in[2]:

- John valiantly cut at the frozen loaf, but his knife was too dull to make a dent in it.

This kind of behavior is unusual for "break" verbs, because 'cut' is a change-of-state verb that describes series of actions, while "break" verbs only specify the resulting state of action.

2.1.2 FrameNet

FrameNet proposes roles that lie on the spectrum of generality somewhere between extremely specific roles (like in our airport example) and extremely general roles (like ProtoAgent and ProtoPatient)[1]. The basic concept is that of the ***frame***. A frame is a schematic representation of situations that involve various participants, props, and other conceptual roles. For example, the frame Probability[3], shown below, is invoked by the semantically related nouns chance, chances, likelihood, odds, probability, significance, and is defined as follows:

- *This frame characterizes the likelihood that a* ***Hypothetical_event*** *will happen as a position on a scale of impossible to inevitable. The likelihood can expressed as numerical* ***Odds*** *or a metaphorical representation of the* ***Position*** *on a scale*

Roles defined by this frame are Hypothetical_event, Odds and Position. With the following interpretation:

- ***HYPOTHETICAL_EVENT*** *The event that is expected to happen with a certain likelihood. He's got a small* ***chance of making it out alive***.

[2]Example taken from [12].

[3]All examples from this section can be FrameNet can be found at `https://framenet.icsi.berkeley.edu/fndrupal/`.

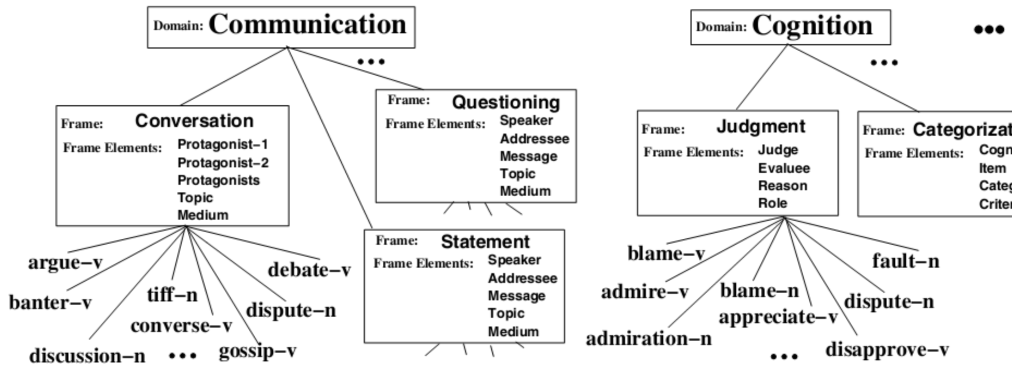


Figure 2.1: Domains of FrameNet defined on varying levels of abstraction.

- ***ODDS*** *A numerical representation of the probability that a Hypothetical_event will occur. There's a* ***95 % chance*** *of rain today.*
- ***POSITION*** *A metaphorical representation of the position on the scale of likelihood that the Hypothetical_event will occur.* ***Chances*** *are* ***slim*** *that he'll pull through.*

The roles are defined on a frame basis and are shared by all of its lexical entries. The diversity of the following example sentences for the Probability frame demonstrate the broad semantic coverage of FrameNet:

- Generally [less]$_{\text{Position}}$ [chance]$_{\text{chance}.n}$ [of temporal variations]$_{H\ \text{event}}$
- [chances]$_{\text{chance}.n}$ are [I attacked the other books too]$_{H_{\text{event}}}$

This annotation clearly shows the level of generality of semantic frames defined by FrameNet. On the one hand, it is specific enough to capture lexical and syntactic information and on the other hand general enough to talk about abstract notions of an inheritance hierarchy of semantic frames. Indeed, FrameNet allows generalizations across different categories of verbs, nouns, and adjectives with each of them adding semantics to the general frame or highlighting a particular aspect of the frame. Conversely, many of the phenomena in the methodology of Framenet remain problematic. For example, it is clear that there is not always a direct correspondence between syntax and semantics. The development methodology of FrameNet should have a big impact on what researchers expect in practical applications. In the first step, a set of semantic frames was chosen for the general domains of interests (see Figure 2.1)[4]. Consequently, a list of target words was compiled for each frame and example sentences were chooses by searching the list of candidates in British National Corpus. Various patterns over lexical items and part-of-speech sequences in the target words' context were performed and the example for annotation chosen with the aim of coverage. Finally, sentences were manually annotated and checked for consistency. It is clear that such an approach emphasizes completeness of examples for lexicographic needs rather real word distribution of semantic phenomena.

[4]Figure taken from [3]

Table 2.1
Subtypes of the ArgM modifier tag

LOC: location	*CAU: cause*
EXT: extent	*TMP: time*
DIS: discourse connectives	*PNC: purpose*
ADV: general-purpose	*MNR: manner*
NEG: negation marker	*DIR: direction*
MOD: modal verb	

Table 2.1: Adjunct semantic roles defined by the PropBank.

Frameset accept.01 "take willingly"

Arg0: Acceptor
Arg1: Thing accepted
Arg2: Accepted-from
Arg3: Attribute

[He]$_{\mathrm{Arg0}}$ [would]$_{\mathrm{ArgM-MOD}}$ [$n't$]$_{\mathrm{ArgM-NEG}}$ accept [anything of value]$_{\mathrm{Arg1}}$ [from them]$_{\mathrm{Arg2}}$

Frameset kick.01 "drive or impel with the foot"

Arg0: Kicker
Arg1: Thing kicked
Arg2: Instrument (defaults to foot)

[John$_i$]$_{\mathrm{Arg0}}$ tried.

Figure 2.2: Sample Framesets as defined by the PropBank.

2.1.3 PropBank

The issues with the broad-coverage and statistically unrepresentative samples of the FrameNet are what the PropBank corpus is trying to tackle. Taking into account that the other end of spectrum (defining a small set of universal roles) is difficult, the roles are annotated on ***a verb per verb basis***[12]. Individual semantic roles of a predicate are numbered starting from 0. Given a particular verb, Arg0 is most probably the argument that exhibits features of a prototypical Agent while Arg1 is a prototypical Patient or Theme. Further, developers point out that no consistent generalizations can be made across verbs for higher numbered arguments. Claims go further to that the effort was made to define roles consistent with respect to the roles across members of VerbNet classes. In addition to these core roles, more general roles that can apply to any verb were defined. The adjunct roles of PropBank are listed in the Table 2.1.

Distinct usages of a verb are captured by the set of its semantic roles, which is called a Roleset. The ***Roleset***can be associated with the set of syntactic frames

Frameset: decline.01 “go down incrementally”

Arg1: entity going down
Arg2: amount gone down by, EXT
Arg3: start point
Arg4: end point

[its net income]$_{\text{Arg1}}$ declining [42]$_{\text{Arg2-EXT}}$
[to \$121million in the first 9 months of 1989]$_{\text{ArgM-TMP}}$.

Frameset: decline.02 “demure, reject”

Arg0: agent
Arg1: rejected thing

[A spokesman]$_{\text{Arg0}}$ declined [$*$trace $*$ to elaborate]$_{\text{Arg2-EXT}}$

Figure 2.3: Defining verb meaning by the number of verb’s arguments.

Frameset open.01 “cause to open”

Arg0: agent
Arg1: thing opened
Arg2: instrument

[John]$_{\text{Arg0}}$ opened [the door]$_{\text{Arg1}}$

[The door]$_{\text{Arg0}}$ opened

[John]$_{\text{Arg0}}$ opened [the door]$_{\text{Arg1}}$ [with his foot]$_{\text{Arg2}}$

Figure 2.4: Sentences with transitive and intransitive usege of the verb “open”.

that suggest allowable syntactic variations with respect to that set of roles and jointly they constitute a Frameset. Consequently, polysemous verbs may have more than one Frameset as represented by the defined differences in meaning. Figure 2.2 show sample Framesets[5].

Discriminative criteria for distinguish framesets are based on both syntax and semantics. For example, two verb meanings are different if they take different number of arguments (See Figure 2.3).[5]

Furthermore, verbs which do preserve the meaning with an alternation such as causative/inchoative or object deletion belong to the same frameset, while allowing for the case in which some arguments could be left unspecified. Such are the examples for transitive and intransitive uses of the verb “open” as dipicted in Figure 2.4.[5]

[5] Example taken from [12]

Frameset see.01 "view"
Arg0: viewer Arg1: thing viewed
[John]$_{\text{Arg0}}$ saw [the President]$_{\text{Arg1}}$ [John]$_{\text{Arg0}}$ saw [the President collapse]$_{\text{Arg1}}$

Figure 2.5: Example of an syntactic misleading for FrameSet indentification.

Finally, as with any system of rules, the syntactic type of the arguments does not constitute the criterion for distinguishing between framesets where both a NP object or a clause object satisfy the constrains(i.e. See Figure 2.5).[5]

3. Computational modeling of semantics

3.1 Supervised learning

Supervised semantic parsing has been usually captured with the following subtasks:

- identifying the boundaries of the arguments of the verb predicate and
- labeling them with semantic roles.

As the arguments can be continuous or discontinuous sequences of words, any subsequence of words in a sentence is an argument candidate. The argument identification has been usually tackled with heuristics or by training discriminative classifiers for prediction. The following task then takes argument candidates and using feature-rich classifiers assigns semantic labeling to them.[5][9]. We depict this standard pipilene in Figure 3.1. The state-of-the-art system[2] on Chinese, Czech, English and German uses a pipeline of independent, local classifiers that identify the predicate sense, the arguments of the predicates, and the argument labels.

The model generates with a beam search a set of candidates which are then re-ranked using a joint learning approach that combines local models and propositional features. Furthermore, feature selection was done, which improved the performance. A full specification and the description of the state-of-the-art systems would exceed the the scope of this master thesis; we will just comment on the general architecture, drawback and the complexity of the approach. Further, in Section 5 we will argue about the shortcomings of this method. The reader interested in details is encouraged to read some of the state-of-the-art research like [19][20]. From another perspective, previous work has also shown good usage of the given architecture and even employ structural and linguistic constraints into the semantic parsing problem. Thus, Punyakanok et al.[15] tackle a problem with

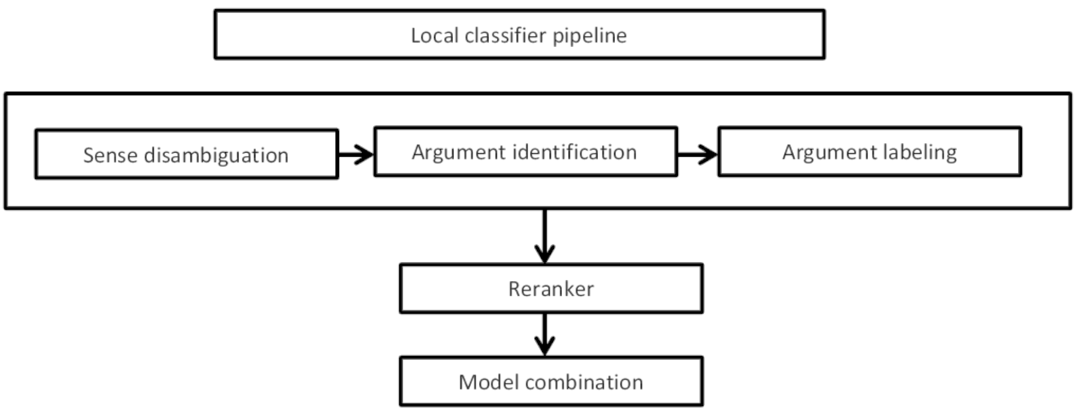


Figure 3.1: The standard supervised architecture for semantic role labelling.

the following setup: given a predicate, they treat all possible spans as candidate arguments and at the first stage pruning, which is done using a syntactic parse tree, they deterministically filter the space. This is followed by further filtering using an argument identification classifier and an argument classification classifier which assigns labels to candidates (both are trained on observable features from the input). In the final stage, all labeled arguments with their posterior probability and a set of linguistically and structurally motivated constraints is submitted to the ILP in order to make a globally consistent prediction. This particular task is well suited for incorporating constraints and at the same time extremely hard to accomplish with global models. Constraints follow structural or linguistic properties: arguments cannot overlap with the predicate; arguments cannot exclusively overlap with the clauses; if a predicate is outside a clause, its arguments cannot be embedded in that clause and many more. See [15] for the formal description of the problem, features used and other details.

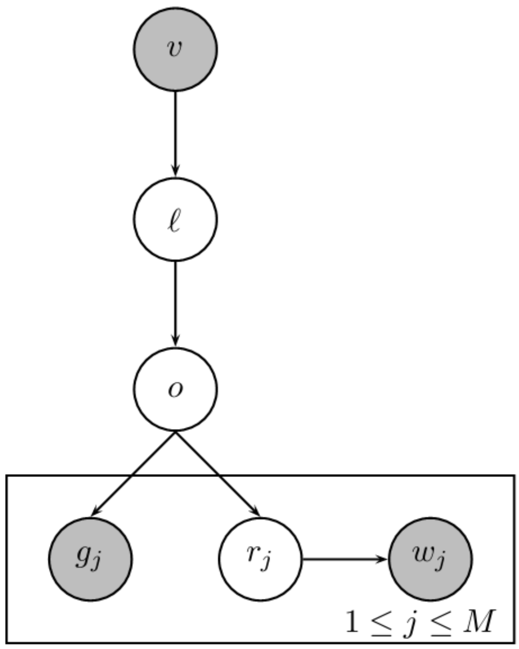


Figure 3.2: Graphical model of Grenager and Manning for unsupervised semantic role labelling.

3.2 Unsupervised learning

Grenager and Manning[4] presented one of the first fully unsupervised systems for semantic role labeling with the aim a broad-coverage language lexical resource. The model was aimed to learn valuable statistical verb behavior that can be easily extended to new text genres and languages. Specifically, the model relates a verb, its semantic verb and their possible syntactic alternations. Syntax was not modeled but gained from corpora annotation or automatic parsers and translated into a fairly language-independent set of syntactic relations, a subset form of a dependency grammar. Furthermore, a simplistic representation of semantic was devised which only had five core arguments, similar to PropBank and one adjunct role which was shared by all the verbs. Table 3.2 offers an illustration of extracted syntactic and semantic representation..

A deeper market plunge today
could give them thier first test.

Verb: give

Syntactic Relation	Semantic Role	Head Word
subj	ARG0	plunge/NN
np	ARGM	today/NN
np#1	ARG2	they/PRP
np#2	ARG1	test/NN

$$v = \text{give}$$

$$l = \{\text{ARG0} \rightarrow \text{subj}, \text{ARG1} \rightarrow \text{np\#2}, \text{ARG2} \rightarrow \text{np\#1}\}$$
$$o = [(\text{ARG0}, \text{subj}), (\text{ARGM,?}), (\text{ARG1,np\#1}), (\text{ARG2,np\#2})].$$

$$(g_1, r_1, w_1) = (\text{subj}, \text{ARG0}, \text{plunge/NN})$$
$$(g_2, r_2, w_2) = (\text{np}, \text{ARGM}, \text{today/NN})$$
$$(g_3, r_3, w_3) = (\text{np1}, \text{ARG\#2}, \text{they/PRP})$$
$$(g_4, r_4, w_4) = (\text{np2}, \text{ARG\#1}, \text{test/NN})$$

Table 3.2: Example of extracting syntactic and semantic representation by the model.

The graphical representation of the model is given in Figure 3.2. The model defines a joint probability distribution over elements of a single verb instance: verb type, semantic role and the head word. The model first generates a verb and then, conditioned on the choice of the verb, it generates the linking which in turn defines a set of core semantic roles and the syntactic relations that express them. One possible drawback with this kind of a model is that the linking is specified only for core semantic roles and the process introduces uncertainty about the choice of linking and its representation in the ordered list. Consequently, an additional variable had to be introduced in order to capture this uncertainty, which in fact increased the complexity of the model. Further recent work has also shown that clustering predicates can be beneficial to the task at question. Titov and Klementiev [17] have explored this kind of an approach while unsupervisely learning semantics for the task of question answering.

4. Probablistic modeling of uncertanty

4.1 Latent Probabilistic Context Free Grammars

Latent probabilistic context–free grammar (LPCF) is a ***generative*** model of parse trees. The observed variables correspond to parse trees and are treated as ***incomplete*** data. When each observed variable T gets labeled (clustered) with the latent variable assignment, the resulting variable is $T[X]$ completely observed.

For example consider the pair of observed and unobserved variables in shown in Figure 4.1.

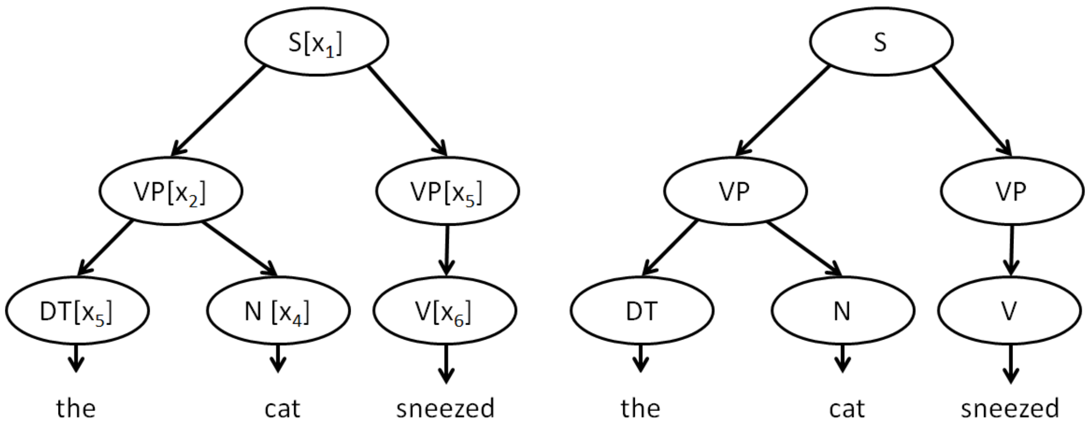


Figure 4.1: Observed and hidden trees of the example sentence.

A grammar that generates complete parse tree is generated exactly as in context free grammar with an the exception that every observed node has to be specified (***clustered***) with a latent annotation symbol.

4.1.1 Model

We use the formulations from [10] and [13]. Formally, $LPCFGLG$ as is a tuple $G = \langle N_{\text{nt}}, N_t, H, R, \pi, \beta \rangle$, where:

- is a set of observable non-terminal symbols
- is a set of terminal symbols
- is a set of latent variable symbols
- is a set of observable CFG rules
- is the probability of the root taking assignment
- is the rule probability

Thus the probability of the complete parse tree from Figure 1 can be computed as:

$P(T[X]) = \pi(S[x_1]) \prod_{r \in D_{t[X]}} \beta(r) = \pi(S[x_1]) \times \beta(S[x_2] \to \mathrm{NP}[x_2]\,\mathrm{VP}[x_5]) \times$
$\beta(\mathrm{NP}[x_2] \to \mathrm{DT}[x_3]\,N[x_4]) \times \beta(\mathrm{DT}[x_c] \to \text{the}) \times \beta(N[x_4] \to \text{cat})$
$\times \beta(\mathrm{VP}[x_5] \to V[x_6]) \times \beta(V[x_6] \to \text{sneezed})$

where , as usually, $P(S[x_1])$ is the probability of generating $S[x_1]$ as the root symbol and the is the probability of the rule . Furthermore, the probability the of the observed tree is gained by summing out the latent annotation symbols X:

$$P(T[X]) = \sum_{x \in H^m} \pi(A_1[x_1]) \prod_{r \in D_{T[X]}} \beta(r) = \sum_{x_1 \in H} \sum_{x_2 \in H} \cdots \sum_{x_n \in H} P(T[X])$$

which has the exponential cost. The reason being that calculation at node n has a cost that exponentially grows with the number of $n's$ daughters because the summation involves $|H|^{d+1}$ combination of latent variables assuming that n had d daughters. However this equation can be computed using dynamic programming methods.

For this purpose, we need definitions of ***forward*** and ***backward*** probabilities in the context of LPCFG.

Thus given a sentence $w_1 w_2 ... w_n$ and its corresponding parse tree T backward probabilities $\beta_T^i(x)$ are computed as:

- If node i is a preterminal node above a terminal symbol w_j:

$$\beta_T^i(x) = \beta(N_i[x] \to w_j)$$

- Otherwise, let j and k be the two daughters of the node i then:

$$\beta_T^i(x) = \sum_{x_j, x_k \in H} \beta(N_i[x] \to N_j[x_j]\,N_k[x_k]) \times \beta_T^j(x_j)\,\beta_T^k(x_k)$$

where i is the nonterminal label of the -th node. Then the probability of an observed tree is:

where $N_i \in N_T$ is the nonterminal label of the i-th node. Then the probability of an observed tree is:

$$P(T) = \sum_{x_k \in H} \pi(N[x_1])\,\beta_T^i(x_1)$$

And similarly the forward probabilities $\beta_T^i(x_1)$

- If node i is a root node:

$$f_T^i(x) = \pi(N[x_1])$$

- Otherwise, let be the right sibling of i and k its mother:

$$f_T^i(x) = \sum_{x_j, x_k \in H} \beta(N_j[x_j] \to N_i[x_j] N_k[x_k]) \times f_T^j(x_j) \beta_T^k(x_k)$$

Provided annotated corpora of observable trees $T = \{T_1, T_2, ..., T_k\}$ we can estimate the parameters with ***EM algorithm***. The algorithm specifies the values of parameters $\theta = (\beta, \pi)$. Derivation of the EM is similar as for other latent variable models and it is defined as a constraint optimization problem:

$$Q(\theta'|\theta) = \sum_{T_i \in T} \sum_{x_i \in H^{m_i}} P_\theta(X_i|T_i) \log P_{\theta'}(T_i[X_i])$$

which iteratively updates the values of the parameters θ and θ' respectively; $P(X|T) = P(T[X])/P(T)$ is a conditional probability of latent annotations given the an observed tree T.

Given learned parameters θ labeling new sentence w can be formulated as:

$$T_{\text{best}} = \underset{T \in C(w)}{\text{argmax}} P(T|w) = \underset{T \in C(w)}{\text{argmax}} P(T)$$

where is a set of all possible parses of w under observable grammar. The expression above involves so called sum-of-product calculation which can be proved ***intractable*** (NP-hard) for latent variable models. There are few approximations over posterior marginal of the parse tree distribution but here we present ***Max-Rule-Product*** objective which is one of the most used one in context of LPCFG:

$$T_G = \underset{T}{\text{argmax}} \prod_{e \in T} q(e); q(A \to B, C, i, k, j) = \frac{r(A \to B, C, i, k, j)}{P_{\text{IN}}(\text{root}, 0, n)}$$

and its rule score:

$$r(A \to B, C, i, k, j) = \sum_x \sum_y \sum_z P_{\text{OUT}}(A_x, i, j) P_{\text{IN}}(B_y, i, k) P_{\text{IN}}(C_z, k, j)$$

where A, B, C are nonterminals, x, y, z are latent annotation symbols and i, j, k are between word indices.

When deciding on the number of latent annotation symbols, one usually uses a fixed number of symbols for each variable. However, that approach has been shown to lead to oversplitting and thus faster overfitting of training data and unmanageable growth of the grammar. Petrov et al. tackles this problem by incorporating a split-merge formulation. Specifically, all latent variables are first split in two and then the ones that gave the highest improvement in likelihood are kept, while others are merged back to the state of the previous iteration. For the reader interested in the details, we suggest Petrov et al.[13].

5. Modeling Semantics: Probabilistic latent variable approach

In making a model for computational processing of linguistic structures, one has to delve deeper into the specificalities of the underlying problem. Current approaches to semantic parsing, at least in the domain of Propbank, have largely ignored that important question. As we discussed in Section 3, supervised approaches rely to a huge extent on the following factors:

- availability of sufficient amount of annotated data
- existence of a well-defined set of features relevant to the task
- assumptions about the correctness of the underlying linguistic structures

5.1 Key insights

In what follows, we will question such an approach, emphasizing its strengths and weaknesses. First of all, the availability of a sufficient amount of annotated data is true only for some languages. And in that case, the amount of data required to make appropriate generalizations might not be sufficient and its sufficiency is hard to bound using the current theories. Furthermore, even if such bounds existed, the cost of creating the additional annotated data might be very high. Further, for most of the languages even low-level annotated data on the level of POS tags and syntactic trees are not available. Considering further annotation on the semantic level becomes meaningless when taking into account those simple facts.

The existence of a well defined set of features for a practical task in language processing is a common assumption. It is well known that features extracted from syntactic trees are extremely helpful in semantic parsing. However, for some languages they just might not exist or be very hard to devise. Imagine a researcher proficient in English devising a set of features for cross-lingual semantic parsing between English and Chinese. Or even English and 63 other languages (the number currently supported by an online translation system); this surely seems like an impossible task. From another point of view, current state-of-the-art approaches use millions of features that can be seen as carefully planned trough fit of an algorithm with respect to true hypothesis in terms of domain specificity (lexical features) and structured specificity (linguistic structures specific to a domain or a particular linguistic theory). Current approaches devise a highly varying discriminative function that maps X to Y and nowhere the model was planned to optimize its structure and parameters as an intermediate representation between layers of linguistic information. The latter is a constraint that is imposed by the whole research community in Computational Linguistics. Namely, when it became clear that humans will most probably never be able to capture abstractions that exist in natural language with a set of rules, the field started to employ statistical methods that can shallowly reason about the specified linguistic structures. In that sense, it made a breakthrough in applications and theories that emerged as a consequence of clear empirical evaluation but one striking assumption slipped through. That is, the reasoning was still about the same human derived abstractions, just in a more empirical way. Thus the reasoning is constrained by the structures, layering and other assumptions of the underlying linguistic theory. To cope with these constraints, it is clear that a discriminative approach is a better approximation. That kind of approach has been shown marginally successful in the predictions of the layers of imposed structures, by using lower layers as features. However, not a single real-world task has been resolved. For example, current research in machine translation is very far from being being linguistically. Out current discussion provides a favorable viewpoint for the supporters of unsupervised learning. If the structures are unrepresentative and inherently hard to model, the algorithm that learns them directly from data is a reasonable counter-part. But then one remembers that we are dealing with the most abstract natural phenomena, in which even the most simple possible tasks can be seen as AI-complete. Consequently, successes in fully unsupervised methods are quite questionable and hard to interpret. For example, if you learn constituent-like structures over strings of words should you be evaluated against human-driven structures or against some different form? As it has been shown, the former evaluation criterion is very unfavorable toward unsupervised algorithms even in tasks like POS tagging, where they perform much lower that the supervised approaches. However, if one learns in an unsupervised manner and then uses the learnt structures for some other task, the performance is quite promising. For example [11] shows that by treating dependency structures as completely unobserved and optimizing them to the task of semantic parsing, one can get results in semantic parsing comparable to the approach that is using gold-standard, or predicted by a supervised algorithm, dependency structures. Further, [17] provides an example of the good use of semantic roles learned in an unsupervised manner on the task of biomedial question answering. However, even these unsupervised tasks share the assumption that the other levels of linguistic structure are provided as input to the learning process. Thus the unsupervised learning without any linguistic structures that has a goal to be as predicative for some layer of the linguistic analysis is also doomed to fail. Simply, the loss of information even though being human incomplete interpretation of the language it still has enough interleaved connection to one another. Thus, as we discuss immediately below, we believe that we should use a semi-supervised approach.

5.2 Semantic Role Labeling – semi-supervised approach?

It is clear that purely supervised or purely unsupervised approaches are insufficient for modeling linguistic structures; we will therefore try to devise a semi-supervised approach. In what follows ,we examine what is incomplete or obscure in semantic parsing, what should be treated as observed, what incomplete and what unobserved. Our hypothesized beliefs are driven how by theoretical underpinnings of underlying theories, previous works describing empirical properties also general descriptions about some of the annotated resources (in English particularity). Thus we do not argue by empirically examining data properties as we are trying to tackle problem in a language independent way. Further one cannot prove or disprove our beliefs by themselves as they are to be learned with latent variables which can automatically adapt to the good or the properties of the same. What are model learns, as we show in results section can be taken as a indicative of the claims.

5.2.1 Learn verb classes?

By verb classes, we mean Levin verb classes, as Propbank annotations are built on them. It is clear that we need a level of abstraction among predicates, from the point of view of dealing with sparsity in natural language and from the point of view that mere semantic decompositions exhibit hierarchical structure. When talking about sparsity, the 1M word WSJ of the Penn Treebank is insufficient in quantity and domain coverage to provide many valuable interpretations. For example, a verb like flap occurs only twice across all inflectional forms, which follows that one cannot learn basic alternation patterns from this data alone. However, abstracting away by grouping similar verbs together with respect to some criteria is surely a way to handle this problem. One can take the intuition for clustering predicates from FrameNet, where everything is organized into one big hierarchy. Furthermore, from a simple computational perspective, when something is infrequent one should smooth it using something that is more frequent. Even from a purely linguistic point of view, it is hypothesized that language is exhibiting that kind of an abstraction. Taking into account that FrameNet abstractions are driven by humans and also incomplete and not statistically representative, we ignore the possibility for driving learning with them. Further, if we wanted to smooth, statistically speaking we would have to first cluster our predicates with respect to discriminative criteria, which is actually a good option but more a quantitative one that a linguistically motivated one. Thus our learning objective will try to learn verb classes as represented by the linguistic theory of linking. It is a natural way to follow from many perspectives. Apart from the already mentioned arguments, the PropBank annotation style, the work of Granneger and Manning also showed promising results. Their model, however, was fully unsupervised, while we opt to do semi-supervised learning and so we will use a different formulation that theirs.

5.2.2 Learn linking?

Learning linkings is the main evidence to support the intuition behind linking theory. However, several questions arise: should the linkings be learned so that they are shared across predicates in verb classes? Should we constrain them to be hard-clustered or soft-clustered? The clear fact is that learning linking alternations across verbs is de-facto imposed by our learning objective, but we further aim to this in a form of probabilistic reasoning. The more evidence we get that some verb should inherit alternations from its corresponding verb classes, the more specified the alternations will be, and vice versa.

5.2.3 Learn cross–class roles?

The ProbBank annotation guide clearly disclaims that argument fillers can be seen as shared across different predicates. However, the hope is the indication that the effort was made to do that. One can easily find a pairs of predicates for which some arguments have very similar, if not identical, syntactic and lexical elaboration. For example, in many PropBank sentences the ProtoAgent A0 is elaborated in the exactly the same syntactic and lexical way. However, some other roles, like the ProtoPatient A1 are quite differently elaborated. Further, this kind of similarity or dissimilarity is interleaved between the roles of the samy type across the whole corpus. So the hope is that one can learn it when it is beneficial and neglect it when it is misleading.

5.2.4 Latent roles?

PropBank defines roles that are neither too general nor too coarse-grained. However, when the arguments become verb-class-cross-shared the generality straightforwardly increases. Further, our argument about insufficiency of the human-driven abstraction, as applied to verb-classes, applies here as well. The level of generality of semantic roles is the subject of an ongoing debate in the field of Lingustics and will most likely to remain as such. As we saw in Section 2, two widely accepted standards are PropBank and FrameNet. We argue that one should directly reason over the level of granularity of semantic roles as represented and constrained by the linkings and verb-classes. We see the level of granularity as domain-specific as the Airport example and as general as the two Proto roles to be undoubtedly justified and representative, as long as it is constrained by the overall model with having the highest likelihood. That kind of reasoning drives the semantics to be as self-expressing as possible.

5.2.5 Learn cross-class dependencies?

Latent dependencies can be argued to be helpful in the same way as the connection between syntax and semantics is the latent one. In some cases the two map deterministically, but in some other drive falsifiable clues. Learning the latency between them is just like trying to tackle its omitted full form. Further, Johansson and Nugues have shown that a richer set of syntactic dependencies improves semantic processing. Also unsupervised approaches work on a principle of keys that can be simply viewed as enriched syntactic dependencies (i.e. with aspect, position).

5.2.6 Learn word classes?

To model the appropriate level of granularity between semantic roles and its lexical representations, one certainly needs some form of class-based definition. First of all, the lexical sparseness is a ubiquitous problem; Zipf's law holds in all languages and is one of the main problems in language processing. Abstracting to the level of granularity of grammatical categories or any other stochastically derived form is an option toward handling the problem of sparseness. However, we adapt a learning of the same in the space of lemma-driven latent variables between the semantics and lexical information. Such an approach can be motivated by the fact that language should exhibit a form of semantics on the level between the surface form and the frame semantics, clearly consistent with our philosophy of hierarchical representations of semantics on deferent levels.

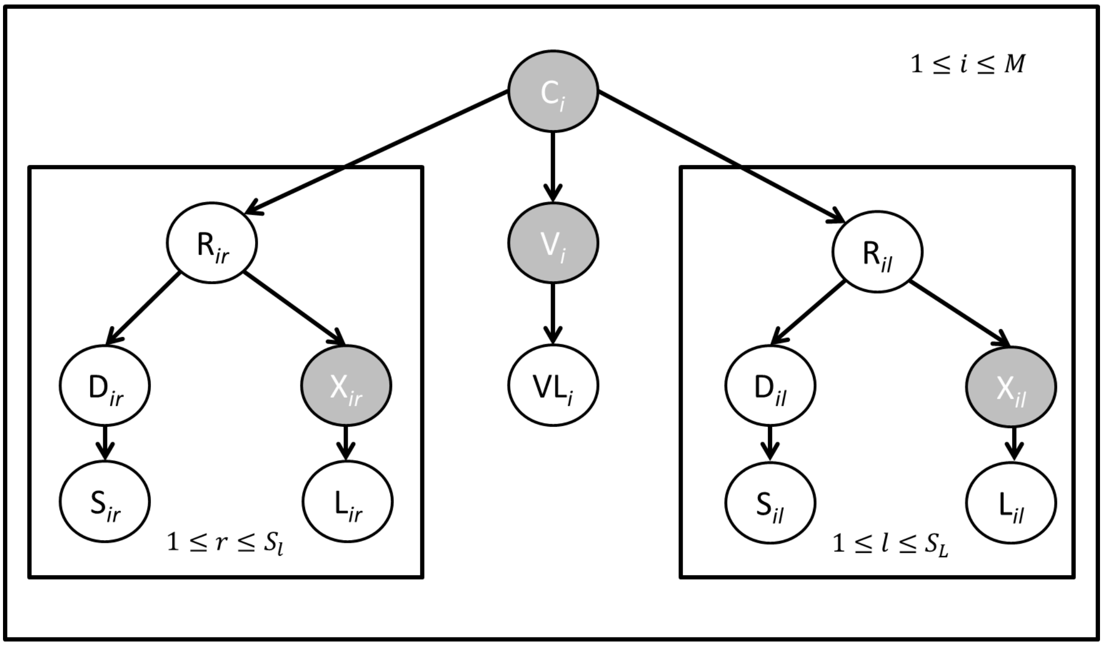
5.2.7 Latent word senses?

The word senses as provided by the PropBank are, to a great extent, already modeled in our approach. It is rather clear that trying to differentiate between numbers of semantic arguments of a predicate will not change anything, since such are already modeled by the learned linkings. Further, the overall compactness of the model will constrain the allowed syntactic and lexical specifications in such a way that they will capture much of the sense meaning.

5.3 Model

In order to accomplish our ultimate goal of learning latent information between many layers of linguistic knowledge, we argue about the modeling perspective in the domain of probabilistic models. Our main goal is to learn varying abstractions of semantics and their corresponding coupling with syntactic and lexical information. Further, our modeling problem has two folds: in one it tries to learn abstractions and generalizations about verb classes, linking, role fillers and word classes and in the second specifications and encapsulation of elaboration of semantics in its lexical and syntactic form. That makes defining a model and its corresponding learning objective quite difficult.

In what follows, we argue that the appropriate level of abstraction and encapsulation can be found in close correspondence of semantic and syntactic dependencies. As it has been shown by [6][7], there is a very high correlation of mapping between syntactic and semantic dependencies. Furthermore, in their approach, which is formulated as an unsupervised learning problem, lexical information plays a crucial role in the unsupervised discovery of semantic role fillers. Granger and Manning's approach is also using a minimal level of interaction between syntax and semantics, as shown in section 3. A further recent success is unsupervised semantic parsing by cross-verb clustering, where the syntactic dependencies are the main information used[14][16]. We thus argue that one does not need to fully specify the complete derivation of the semantic elaboration, and the particular aspects of syntactic and lexical information can be found in minimum correspondence.The graphical model representation of our model is given in the Figure 5.1.



Obrázek 5.1: Latent variable approach for semi–supervised semantic parsing.

The model specifies incorporates intuition about of linguistic theory and the probabilistic modeling in the following ways. The generative story goes in intuitive and simple direction: First one generates a frame class – C, its corresponding linking and the verb class – V. The linking is not fully specified by the frame class variable but it constrains probabilistically at the current moment at least in the number of arguments. The verb class variable and the frame class variable we adapt for explanation purposes, as they are in fact two very closely related variables and they specify complementary information. If we would like to be as ambitious we would say that the frame class should be predicative of the frame, as in the sense of FrameNet at some level of abstraction, while the verb class variable would then in fact specify a particular event or property. Further for each argument filer the semantic role is generated – R, which in turn generates its syntactic – D and lexical elaboration –X. Syntax is represented by the dependency arcs between the semantic role and its governor, while the lexical information is generated through a word class. Word class in fact is not to be taken as in its usual interpretation of the entity that groups similar words from a linear sequence as based on their context. Rather here the word class captures the interaction between a semantic role and its possible lexical information. We use the lemma of the surface form to represent the lexical information.

As our model is currently specified, it is a very simple model indeed. One could easily argue that it cannot capture many phenomena in natural language that are influenced by the same type of linguistic knowledge that we are using. However our goal is not to fully specify all the forms of linguistic knowledge that we are using but rather only one: the semantics.

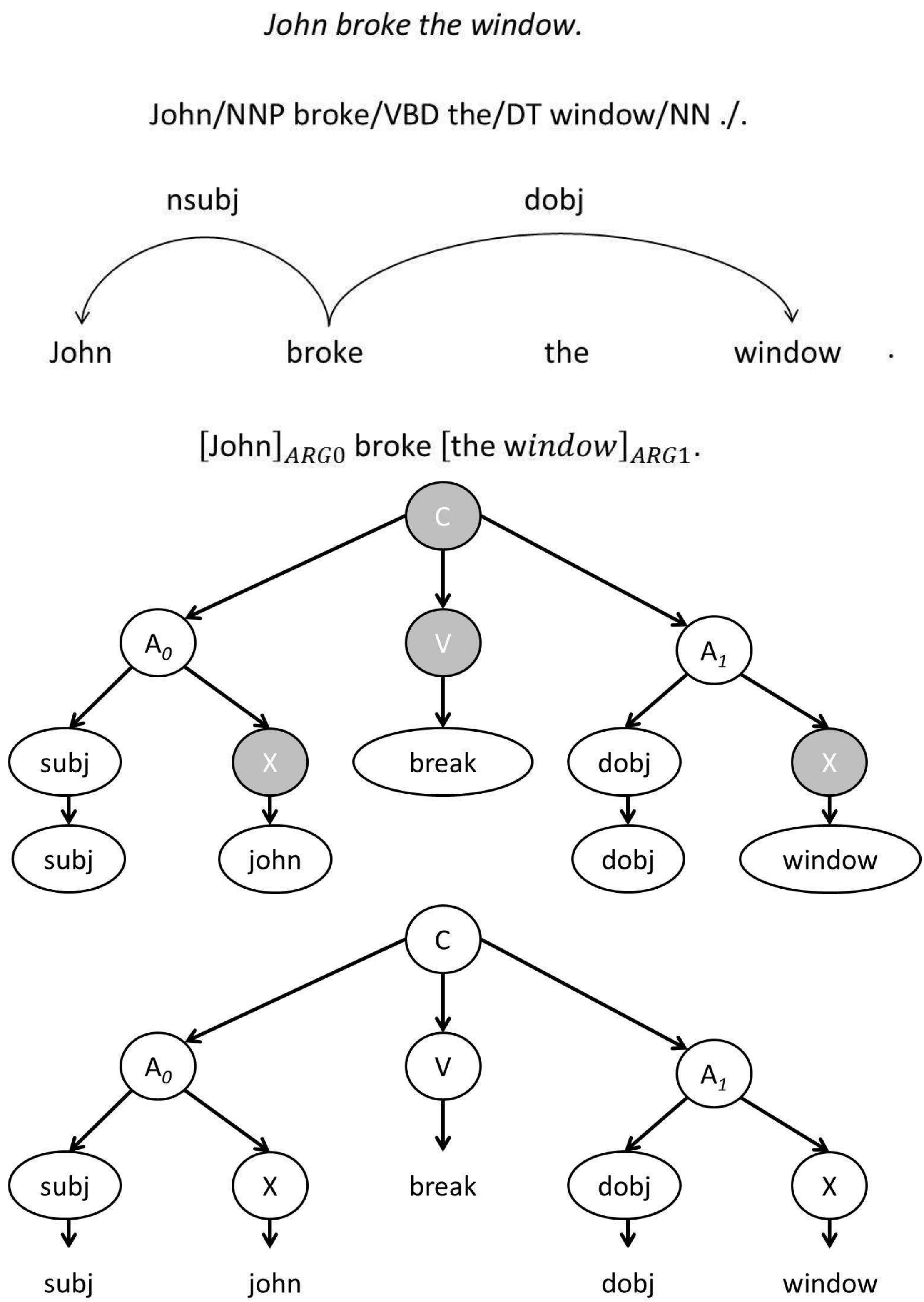
Thus our model should be only predicative of the roles given all other observable arguments. Further, even in this very simple model we have three unobservable types of variables. We do not get to observe the frame-class variable, the verb-class variable and the word-class variable. In the simplistic type of model as the one on the Figure 5.1 , that would limit the expected performance of the model in a high degree. Consequently, we tackle the described problems as well as the full motivation behind the modeling by adapting latent variables on each node. Then the non-observability of the variables in our model actually becomes its expressive power. Further, all of our model variables are shared across different semantic frame instances, on the sentence level as well as on the corpus level. Thus the model will be able to learn varying degrees of semantic knowledge as represented by all: frame-classes, verb-classes, dependencies, word-classes and the linking. Also note that now, since we do not observe the class variable, the arguments variables loose their independence assumptions. Further, from a probabilistic point of view, the model is very compact thus the correlation between variables should be stronger and its learning easier. Most importantly, our model does not use any features so the model is language-independent; no changes in the model are required to handle new language instances.

5.4 From the modeling to reality

Our graphical model could be realized in many different forms of probabilistic learning and inference. First let us consider what the current model might have the problem capturing. As it is represented the model does not incorporate any prior knowledge on any type of variables. For sure that kind of information is very useful in reasoning over linguistic knowledge. Many variables, if not all, from our model could valuably incorporate the priors. However, one of them has been shown very crucial in dealing with semantic parsing. The linking prior which is the main component of the Graneger and Maning's model has been used in semantic parsing since its pioneered work in the task [3] till the state of the art models in unsupervised parsing of the state–of–the–art research[18]. Further even though our model is compact in dealing with semantic parsing with cannot guarantee or assume the number of hidden variables. That implies that some form of split-merge adaptation that can automatically adapt at the new data should be used. Further as for all latent variables models the intractability imposed the need for well tackled and implemented approximate inference algorithms. All sad, we decide to use latent variable probababilistic context free grammar as the formalism for the model realization. With observing our model as a LPCFGs we naturally capture priors over structures as the same are defined over context-free rules and in that way tackle learning of linking compactly in our formulation. Furher, LPCFGs have been so far very successfully used in problems of syntactic parsing and have developed advanced learning and inference procedures. One of them is developed in the Berkley parser implementation of LPCFG. Thus we only need to formulate our problem in terms of parsing with context-free grammar and we use the Berkeley parser[1] as the of-the-shelf tool. This conversion is straightforward and we depict it in Figure 5.2.

The formal form of the model follows LPCFG, as we observe each semantic frame as a tree in the context-free grammar form. Thus the mathematical underpinnings are already defined and the reader is encouraged to see section 3 for related references. One can easily see many alternations of our model as depicted by some linguistic property. For example, we could instead of word class X observe some word class from an external clustering (i.e. Brown classes). The verb could be represented by its lemma or surface form as well as the arguments' lexical heads. Further, one could incorporate POS information on arguments as well as predicates.

[1]http://berkeleyparser.googlecode.com/files/BerkeleyParser.jar



Obrázek 5.2: Intuitive casting process of our model to LPCFG. We are given with the desctiption data annotated by the model from Figure 5.1. Thus implies that the required levels of analysis on the lower level of linguistic structures have been done. Then we simply map the variables to the PCFG trees.

6. Empirical plausibility

6.1 CONLL09

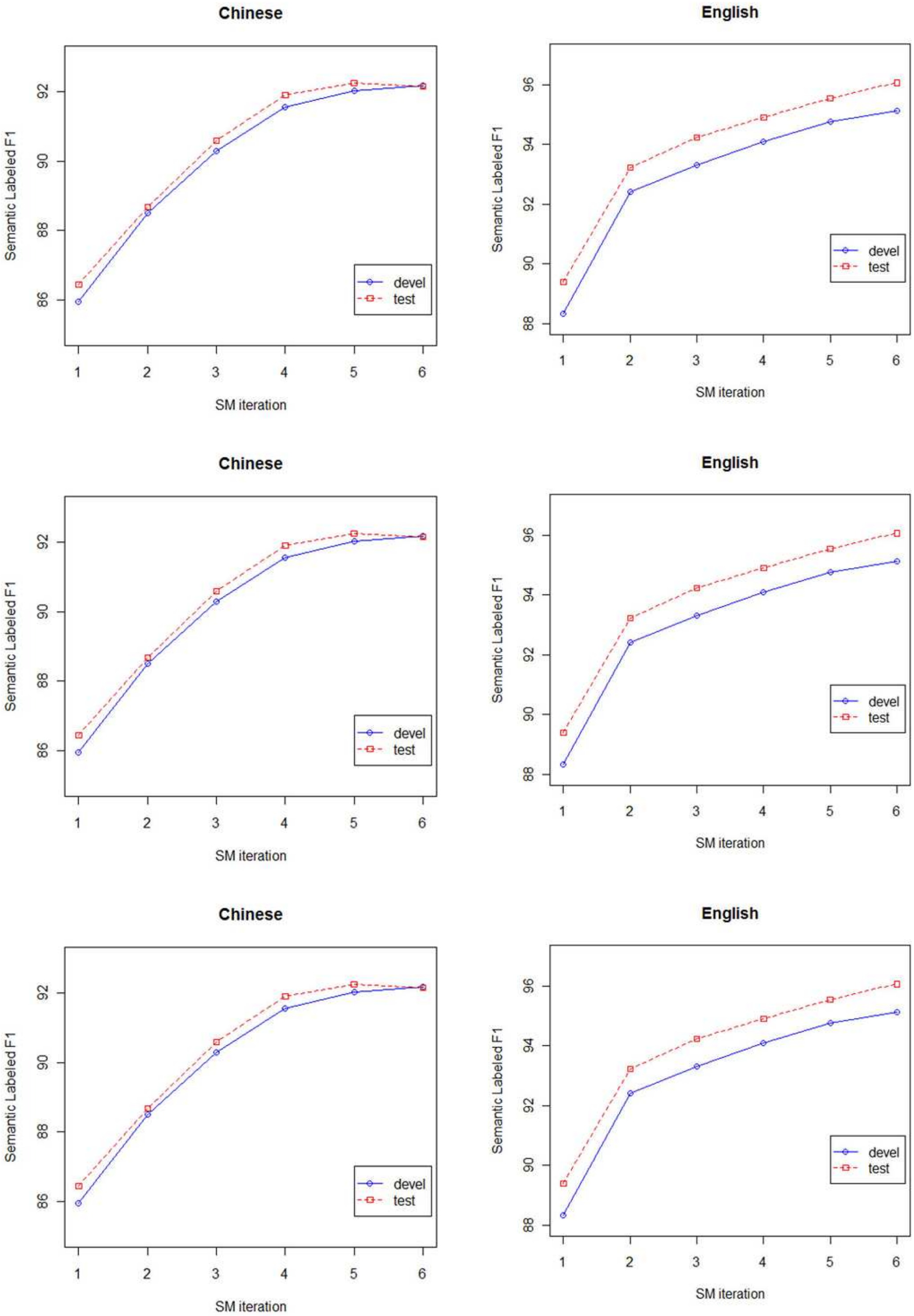
We evaluate our model on the data provided by the CoNLL-09 shared task[5]. The focus of the task was to perform joint learning and inference over syntactic and semantic dependency structures. The semantic relations included, apart from the standard verbal predicates, propositions over other major part-of-speech categories. However, in our current investigation we implicitly assumed that we are dealing with verbal predicates and in order to learn varying degrees of linguistic structures we train and evaluate only over verbal predicates. Our argument is that even though that information provided by other propositions might be complementary to our final task of interest, its motivation in the underlying theories which we are using is not clearly defined, most certainly not cross-linguistically. For the details of the data and the task we encourage the reader to see Hajic et al. [5].

From the previous section, is it quite clear how we would do the training of our model. We extract relevant information from the training data, run Berkley parser with some split-merge step, go until our computational resources allow, then evaluate on the development set and choose the best split-merge iteration for using on the test set. However, our model assumes that the arguments are known to us. We leave interaction of the argument identification and learning latent semantics to future work and we focus on being predicative of the semantic propositions. The task of argument identification is solved with an extreme precision, with discriminative methods having a cross-language accuracy of over 95%. Thus when evaluating on unknown argument boundaries we use outputs of the near state-of-the-art system Nugues [2]. Further, we use the outputs of the system on all other lower level structures: POS tags, dependencies and the labeling of non-verbal predicates. We decided to do so as the Nugues system is freely available, so one could use the combination of our approach and their system in the real-word application directly. Further, that gives us a clear empirical setting and the possibility of using benchmark evaluation provided by the task (eval09.pl script[1]).

6.1.1 Results

Figure 6.1 the performance of our system on development and test set with varying split-merge steps on the known arguments. We only show the performance on the test set for explanatory purposes and insights about the upper bound on the performance which we can expect when combined with argument identification task. Further, one should note that our results on known arguments are not to be taken as directly comparable with other approaches as we never predict non-verbal propositions. In the case of the testing on known arguments they get exact match on evaluation as we are using gold standard data as the input to the system. Further, on the task when arguments are not known we use Nugues system's non-verbal proposition predictions. We decided to do evaluation in that way so that we can use benchmark evaluation provided by the task. Non-verbal

[1] `http://ufal.mff.cuni.cz/conll2009-st/eval09.pl`



Obrázek 6.1: Performance of the model with respect to split–merge iterations.

Rank	System	Average	Catalan	Chinese	Czech	English	German	Japanese	Spanish
1	Zhao	80.47	80.32	77.72	85.19	85.44	75.99	78.15	80.46
2	Nugues	80.31	80.01	78.60	85.41	85.63	79.71	76.3	76.52
3	Meza-Ruiz	77.46	78	77.73	75.75	83.34	73.52	76	77.91
	Our work	**76.46**	**76.23**	**75.11**	**79.82**	**83.12**	**73.50**	**72.17**	**75.28**
4	Baoli Li	69.26	74.06	70.37	57.46	69.63	67.76	72.03	73.54
5	Moreau	66.49	65.6	67.37	71.74	72.14	66.5	57.75	64.33
6	Täckström	61.27	57.11	63.41	71.05	67.64	53.42	54.74	61.51
7	Lin	57.18	61.7	70.33	60.43	65.66	59.51	23.78	58.87

Obrázek 6.2: Final performance of our system compared with the CoNLL09 participating systems.

arguments occur in quite lower number of times than the verbal ones. By using predictions from another system we gain certain relaxation in overall score but still quite close to reality. If we have decided to build our own evaluation tool still the question of consistently differentiating verbal vs non-verbal arguments would have to be dealt with, both in training and in inference, which is not a deterministic process and the comparison with other approaches would be hard to interpret. In this way, we use Nugues system's for all except verbal predicates which are totally predicted by our model and thus we can compare our performance with other systems on the same task.

We select the best performing split-merge iteration number for each language based on the best development set performance and run our system on the test. Table 6.2 show final test results of our system and other systems that competed on the CoNLL09 shared task on predicting semantic dependencies.

As you can see our model is performing comparable to best of the systems of the CoNLL09 shared task. Model automatically learns appropriate levels of abstraction needed for good performance. However our model was not optimized for the task but rather it was optimized to be good model of the underlying structures over which is defined as measured on the recovery of training data or training data likelihood if you will. Further our model does not incorporate any prior knowledge about the structures which are relevant in the learning process. The clear fact is that we would like for instance to split more our unobservable variables. To give some insights about the clustering which are model learns from training data we show statistics about most frequent assignments to some of the variables of our model. Further for the purposes of completeness and the clearer interpretation of the results we simply run our model on semantic arguments both verbal and non-verbal. That kind of an approaches clearly skewed the probability distribution of the outcomming model, since the same was designed for the purpose of the predicate modeling. We could for example run our model separately, thus train two models and to inference separately as well for verbal vs non-verbal arguments but the problem of differentiating among the same made as to leave this to future work. We get the result for English 80.35% F1 on 5 split–merge iterations. Thus our model is even performing on the level that is close to the state–of–the–art on the whole semantic prediction task. Surely that shows to full power of our approach.

7. Future work

We have shown promising results of modeling varying abstraction of semantics trained jointly with syntactic and lexical information. Out model has a simple and compact form optimizing only training likelihood. To further increase the expressiveness of our model as well as the performance one would have many options. First of all, the training objective could be changed so that model directly optimizes some form of validation error or validation likelihood. One might also what to consider using some other graphical model implementation and thus delve into the technicalities of the problem. Learning the non-verbal argument representation was neglected in our work and should be definitely be considered in future. Jointly learning full derivation of syntactic and lexical representation of the semantics in a single model is definitely required if the model would be considered for the real-word application. The approach presented is very simple and breaks a lot of independence assumptions of the current approaches to semantic parsing at the same time without using any features while abstracting and encapsulationing required information. That suggests that it might can be used in even more ambitious goal of inducing semi-supervised abstractions of linguistic structure over multiple languages.

Conclusion

One of the reasonable approaches in dealing with language processing, at least while the language is seen as a string of tokens, is to combine linguistic structures and powerful structure learning algorithms. The first being the necessary word knowledge in some of its forms and the second being empirical reasoning over obscure, incomplete and noisy data. That kind of an approach is using linguistic structures but threat them as a backbone structure for learning while trying to specify the structures and parameters in order to perform well on the task of interest. We have presented semi-supervised latent variable approach for learning varying levels of semantics. Our model does not use any features while jointly learning syntactic and semantic dependencies as suggested by the linking theory. Our model in its simple form shows good cross–lingual performance without any changes in the model. Further we have shown quite a radical new approach which ignores verb-per-verb assumption, that learns linking compactly in the model, that assumes role fillers to be cross-shared, dependency structures underspecified and word senses unnecessary. Most importantly we learned semantic frames with varying levels of abstraction. That gives a hope to the aim of semantic parsing of multilingual free text while its direct abstractions are learned in the task based manner.

Literatura

[1] Baker, C. F., Fillmore, C. J., and Lowe, J. B. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1* (Stroudsburg, PA, USA, 1998), COLING '98, Association for Computational Linguistics, pp. 86–90.

[2] Björkelund, A., Hafdell, L., and Nugues, P. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task* (Stroudsburg, PA, USA, 2009), CoNLL '09, Association for Computational Linguistics, pp. 43–48.

[3] Gildea, D., and Jurafsky, D. Automatic labeling of semantic roles. *Computational Linguistics 28* (2001), 245–288.

[4] Grenager, T., and Manning, C. D. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2006), EMNLP '06, Association for Computational Linguistics, pp. 1–8.

[5] Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. The conll-2009 shared task: syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task* (Stroudsburg, PA, USA, 2009), CoNLL '09, Association for Computational Linguistics, pp. 1–18.

[6] Lang, J., and Lapata, M. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Stroudsburg, PA, USA, 2010), HLT '10, Association for Computational Linguistics, pp. 939–947.

[7] Lang, J., and Lapata, M. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2011), EMNLP '11, Association for Computational Linguistics, pp. 1320–1331.

[8] Levin, B. *English Verb Classes And Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, 1993.

[9] Màrquez, L., Carreras, X., Litkowski, K. C., and Stevenson, S. Semantic role labeling: an introduction to the special issue. *Comput. Linguist. 34*, 2 (June 2008), 145–159.

[10] Matsuzaki, T., Miyao, Y., and Tsujii, J. Probabilistic cfg with latent annotations, 2005.

[11] Naradowsky, J., Riedel, S., and Smith, D. A. Improving nlp through marginalization of hidden syntactic structure. In *Proceedings of the Conference on Empirical methods in natural language processing (EMNLP '12)* (Jeju, Korea, July 2012), Association for Computational Linguistics.

[12] Palmer, M., Gildea, D., and Kingsbury, P. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist. 31*, 1 (Mar. 2005), 71–106.

[13] Petrov, S., Barrett, L., Thibaux, R., and Klein, D. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (Sydney, Australia, July 2006), Association for Computational Linguistics, pp. 433–440.

[14] Poon, H., and Domingos, P. Unsupervised Semantic Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, Aug. 2009), Association for Computational Linguistics, pp. 1–10.

[15] Punyakanok, V., Roth, D., and Yih, W.-t. The importance of syntactic parsing and inference in semantic role labeling. *Comput. Linguist. 34*, 2 (June 2008), 257–287.

[16] Titov, I., and Klementiev, A. A bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 2011), Association for Computational Linguistics, pp. 1445–1455.

[17] Titov, I., and Klementiev, A. A bayesian approach to unsupervised semantic role induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France, April 2012).

[18] Titov, I., and Klementiev, A. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (Jeju Island, South Korea, July 2012), Association for Computational Linguistics.

[19] Zhao, H., Chen, W., Kazama, J., Uchimoto, K., and Torisawa, K. Multilingual dependency learning: exploiting rich features for tagging syntactic and semantic dependencies. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task* (Stroudsburg, PA, USA, 2009), CoNLL '09, Association for Computational Linguistics, pp. 61–66.

[20] Zhao, H., Chen, W., Kazama, J., Uchimoto, K., and Torisawa, K. Multilingual dependency learning: Exploiting rich features for tagging syntactic and semantic dependencies. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared*

Task (Boulder, Colorado, June 2009), Association for Computational Linguistics, pp. 61–66.